BLOOR RESEARCH

SPOTLIGHT

Authors
DAVID NORFOLK (LEAD)
PAUL BEVAN (CO-AUTHOR)

Bloor Research
SPOTLIGHT

JULY 2025

Business opportunities from managing and anonymizing healthcare information

Executive summary

This document spotlights a business opportunity for healthcare organizations to improve the service that they give to both clinicians and patients, using the application of advanced analytics to patient healthcare data. Moreover, to do this without putting compliance and data privacy at risk. It is aimed at health service managers generally and, in particular, at technicians and data scientists responsible for managing the access to potentially sensitive, health data. We attempt to take a global view, but we don't have the space to go into the specific details for different countries.

Healthcare data has special characteristics. Not only can it be particularly sensitive (it can wreck people's lives if it gets into the wrong hands); breaches are of particular interest to the Press, TV News etc. It is also extremely regulated by governments. One should never overlook Reputation Risk.

Healthcare data also has special technical characteristics. There are good standards, such as FHIR, but healthcare data tends to stick around for the life of the patient – potentially 100 years or so, so you can easily have to deal with obsolete formats or even hand-written notes scribbled in the margins of documents or forms.

What this all means, in what is usually a risk-averse medical environment, is that data may be collected, expensively managed and stored, but never actually used, because of the perceived risk of so doing. The scope of this paper mainly includes the protection and anonymization of sensitive healthcare (medical) data, so that it can be used productively. We will touch on emerging issues (such as the use of AI, data discovery and quantum computing) and data discovery, as well as legal issues, but these will not be covered in detail. For reference, Daniel Howard at Bloor is in the process of preparing a general Market Update on the data discovery market, for release by Bloor in the summer of 2025.

In summary, the business opportunity we'll explore in this Spotlight paper is the safe utilization of medical data generally, to improve patient care and clinical effectiveness across healthcare systems. This involves securing and anonymizing personal health information in such a way as to free up more data for analysis, without putting patient (and clinician) privacy at risk. Modern technology and tools have made privacy-aware analysis easier than ever before, and you can use this analysis to improve service to customers (and regulators) at comparatively low cost. It is the gift that keeps on giving.

> "The scope of this paper mainly includes the protection and anonymization of sensitive healthcare (medical) data, so that it can be used productively."

The business opportunity in more detail

In brief, the business opportunity is to use health care information, whether private or not, to improve the efficiency and effectiveness of health care systems. The constraint is that privacy must be protected, since if a person's medical history becomes public property, or even if it is just available to a few unauthorized people (in, say, an insurance company or human resources department), that person's welfare may be seriously affected.

The first line of protection is the law, which should say what information can be shared, and who with, and what levels of privacy need to be maintained. One problem is that laws differ in different countries. In the USA, as a result of the 1966 Health Insurance Portability and Accountability Act (HIPAA) compliance is effective and well understood. A data set is considered to be sufficiently anonymized (so that particular individuals cannot be associated with their data) for use in analytics processing, if there is a *"sufficiently small"* risk of de-anonymizing someone (see the HIPAA journal for more detail on this). In Europe, however, the General Data Protection Regulation (GDPR) applies, with special treatment for health data and a greater reliance on maintaining a demonstrable privacy culture, and (probably) a greater emphasis on individual privacy. What is adequate for HIPAA might or might not satisfy GDPR's demands. And HIPAA defines specific healthcare datatypes, while GDPR doesn't.

We want to use Protected Health Information (PHI), any information that must be secured to safeguard a patient's healthcare privacy, to:

1. Improve the effectiveness and efficiency of the patient experience.
2. Make medical information available to doctors in a complete and timely manner
3. Ensure that adequate patient records are kept, to identify existing conditions, allergies, intolerances and so on.
4. Facilitate research into disease demographics, the spread of epidemics and so on.
5. Facilitate the identification of emerging health risks.
6. Automate first level patient diagnosis and triage.
7. Facilitate the interpretation of complex visual information, in X-rays, for example.
8. Facilitate the interpretation of complex information such as that in DNA sequences.

The constraint is that privacy must be protected, since if a person's medical history becomes public property... that person's welfare may be seriously affected.

Challenges to be aware of

Some of the data use-cases identified require complete information that identifies the patient and is protected on a *"need to know"* basis by professional ethics. Other uses need only aggregated data with the patient name obfuscated.

From the start you need to be thinking about exactly what information your application needs and what it can do without. Referential integrity matters. If, for example, an identifying field such as surname is used to link all of a person's operations back to the person operated on (probably not good practice), that surname must obfuscate to the same string wherever it is used as a key root – or there is a good chance that the data won't make sense and will give meaningless analysis results, and will be useless if used for test data.

What is involved in correctly removing identification from a dataset will depend on the data structure, application design, data sensitivity and so on. It also depends on what legal jurisdiction applies (although EU GDPR is becoming a bit of a de facto standard, adopted as the basis for many local data privacy laws). There is a temptation to adopt the highest and most restrictive privacy standards everywhere, which seems safe, but which may impact adversely the effectiveness and efficiency generally of the patent experience for some patients. It is important that you come up with a plan to recognize what health data resources you have and how you will use them effectively to aid your provision of healthcare. Sometimes this will be a direct benefit, to the speed of diagnosis of a serious condition, perhaps; sometimes it will be an indirect benefit, from the sale of (anonymized) patient data to a commercial drug company, perhaps.

If you have already implemented a data privacy culture, for GDPR perhaps, you will find managing PHI safely and effectively much easier. Try not to implement a PHI privacy silo, just for what you see as health-related information– at best, this increases the risk of employees becoming careless; at worst, it provides a back door into your secure silo from elsewhere in the organization.

> "From the start you need to be thinking about exactly what information your application needs and what it can do without. Referential integrity matters."

First, identify what you have

PHI is defined as different things by different sources. Some wrongly define PHI as patient health data (it isn't) whereas others (particularly in the USA) believe it is defined from the 18 HIPAA identifiers (it's not those either). To best explain what is really considered PHI under HIPAA compliance rules, it is necessary to review the definitions section of the Administrative Simplification Regulations starting with health information. According to this section, health information means any information, including genetic information, whether oral or recorded in any form or medium, that:

"Is created or received by a health care provider, health plan, public health authority, employer, life insurer, school or university, or health care clearinghouse; and relates to the past, present, or future physical or mental health or condition of an individual; the provision of health care to an individual; or the past, present, or future payment for the provision of health care to an individual." This an American definition, but I imagine that it is a good starting point internationally. You will need to reference any definitions provided in local regulations, of course.

Then you need to decide where this *"PHI"* might be hiding and how you can identify it. Obviously, you look in any files labelled as healthcare information first. Increasingly, these will be FHIR records (FHIR, Fast Healthcare Interoperability Resources is a standard for healthcare data exchange, published by HL7. Possibly everybody *"should"* be using digital FHIR formats, soon at least, but (referring to the PHI definition above), you could be missing a lot of PHI if FHIR records are all that you look at. Historical data tends to stick around in historical formats (because converting it to FHIR could be expensive) and diagnoses made when I was 5 years old might still be relevant (and need to be kept private) when I was 70 years old.

So, you need to be aware of a wide range of formats and deal with them. For example:

1 There are mature EDI standards for health data exchange in the USA. X12 HIPPA is a legacy American EDI format, but you might meet it in the UK, say – suppose an American tourist has a stroke on a visit to the UK and spends a month in a UK hospital and her doctor needs her medical notes from the USA. X12 processing isn't trivial: X12 and X12 HIPAA are related but still distinct concepts. X12 refers to a set of standards developed by the Accredited Standards Committee (ASC) X12, which is responsible for developing and maintaining electronic data interchange (EDI) standards in the United States. These standards govern the exchange of business documents, such as purchase orders, invoices, and shipping notices, among trading partners. X12 standards are widely used across various industries to facilitate electronic communication and transactions. X12 HIPAA specifically refers to the subset of X12 standards that are mandated for use in healthcare transactions under the HIPAA act but other EDI records might be considered to be PHI under some circumstances, it seems to us.

2 FHIR (Fast Healthcare Interoperability Resources) is a modern interoperability specification from HL7 International designed to be easier to implement, more open, and more extensible than HL7 versions 2.x or 3.x.

3 Health Level Seven (HL7) is a range of global standards for the transfer of clinical and administrative health data between applications that has been superseded by FHIR but you might still meet it, in historical US data, at least, and in Europe (over time, you would expect to see more FHIR and less HL7 as systems are modernized). HL7 standards focus on the application layer, which is *"layer 7"* in the Open Systems Interconnection model. The standards are produced by Health Level Seven International, an international standards organization, and are adopted by other standards issuing bodies such as American National Standards Institute and International Organization for Standardization. There are a range of primary standards that are commonly used across the industry, as well as secondary standards which are less frequently adopted.

4 An emerging, and sensitive, class of healthcare information is that found in, or written on, medical scans.

5 Then don't forget the pesky *"whether oral or recorded in any form or medium"* in the *"PHI"* definition above. Handwritten notes, recordings, typed letters informative stickers, stuck onto patient cards, polaroids, smartphone photos etc. could all be subject to regulation and privacy.

Automation will be critical to effective management and protection of PHI, but can it manage all the formats you might have? It is very likely that all new systems in the UK NHS, say, should be compatible with FHIR, but by when; and for how long are you still likely to meet historical data not in FHIR formats? And how well will your automation handle unstructured information – note *"whether oral or recorded in any form or medium"* highlighted in the definition above – and audio/video files?

Second, decide what you want to do with what you have

Once you have identified your PHI (and remember that this is a continuing effort, not something you only do once), you will need to catalog it (so you can find it again), classify it by sensitivity and value, and decide what level of security needs to be applied. You need to decide whether to encrypt private data (more secure, less usable) or whether you just need to anonymize key fields, so that the subject of the data can't be identified.

Ideally, you should be using platform-based automation for this, which keeps a record of your policy decisions, and the characteristics of your anonymization workload. You will want a platform that can integrate with your existing toolsets, possibly with API interfaces. You should use a tool that can maintain referential integrity.

You should avoid using spreadsheets or word-processed documents for managing this, in favor of something more structured. Of course, some of your PHI may actually be in spreadsheets etc., in which case they should certainly be included in your de-identification processes.

Finally, don't forget to back up your de-identification metadata regularly. You don't want a data breach of sensitive data after, say, a hardware crash because your de-identification process couldn't be reinstated accurately.

Why should you care about anonymizing healthcare data?

Anonymizing PHI better, faster or more completely (and more accurately) than other organisations can, could well give you a competitive edge. You can use more data for analytics, without risk of a data breach, and this should help you to optimize your processes and procedures.

Fundamentally, you care about anonymizing PHI, because this is key to managing risk:

1. The risk that valuable, but sensitive, data is stored but never made use of – the risk of waste.
2. The risk that valuable, but sensitive, data is stolen or read by unauthorized people – the risk of data breach, with associated sanctions. GDPR fines can be huge and HIPAA now has some teeth, with non-compliant companies being fined and subject to lawsuits.
3. The risk that private information is made public and the organization then looks unprofessional – regulation risk.

> "You need to decide whether to encrypt private data... or whether you just need to anonymize key fields, so that the subject of the data can't be identified."

What are the benefits of analyzing healthcare data

You can't manage what you don't know that you have. Healthcare is regarded as a right by many people and healthcare failures have huge political and financial implications. Effective and defensible management of healthcare provision needs data, and often those data are private to individuals.

Take Public Health England (and we're sure that all national healthcare systems have similar issues). It desperately needs good quality data on patient health outcomes. Getting access to sufficient properly anonymized data is an ongoing challenge. But with such data, there are significant opportunities for better mapping health outcomes, understanding health trends and making better long-term decisions about the way health and social care is provided.

For an individual healthcare facility, such as a hospital, access to data helps it to become a better managed, more stable, more respected organization, that makes fewer mistakes. These benefits go beyond just mitigating risk – better management promotes better morale, and employees work better.

Not that mitigating risk is not to be discounted. Being able to sleep at night is good; as is avoiding fines for data breaches and avoiding the associated disruption; and mitigating reputation risk (by keeping out of the papers, perhaps).

What anonymization features are important in a healthcare analytics solution:

1 It must de-identify both test and production data. This is partly because the 1996 US Health Insurance Portability and Accountability Act (HIPAA) in the USA doesn't distinguish between test and production environments and requires the de-identification of 18 unique patient *"key identifiers"* for its Safe Harbor Security Rule. Mainly, however, because test data is often less well looked after than production data, so that it is an obvious back door route in to stealing PHI.

2 It must offer deterministic data masking functions such as format-preserving encryption or unique, consistent pseudonym replacement values, that can, if necessary, preserve referential integrity in masked environments, for structured, semi-structured, and unstructured datasets.

3 It must support the HIPPA Expert Determination Method security rule, which specifies that datasets may not be more than 20% likely to re-identify a particular individual. To comply with this rule, re-ID risk determination must be statistically measured using approved algorithms like l-diversity or k-anonymity. Some sort of workbench will help users manage this and ensure that the specifications are met. Of course, in many jurisdictions, HIPPA will not apply, but determining an acceptable risk of de-identification (20% seems a bit high to us) will still be useful.

A rich set of additional anonymization functions should be available if needed. For example, blurring (random noise) for dates of birth or treatment, and binning to put quasi-identifiers like diagnosis, drug, profession, location, or marital/education status into a broader bucket. Some anonymization is not easily reversed, which is good of you are selling test data to a third party, but useless if you are sending it over an insecure line – when you'd need strong encryption.

4 It needs to deal with the data needs of outsourced services and vendor collaboration, without compromising patient privacy. This means that it must cope with a wide range of data formats.

5 Data anonymization for the use of data in training and education should be supported – the datasets must be realistic after anonymization but privacy cannot be compromised.

6 Cloud and hybrid cloud environments may merit special attention. Data must be protected in transit and at rest, possibly using a reversible anonymization such as encryption (with some of the anonymization methods, you won't be able to extract the original data from the anonymized dataset).

Architecturally, you probably want an anonymization platform, supporting a wide range of tools and data formats, rather than a single point-to-point anonymization utility, which may be less flexible, less future-proof and less able to scale.

Other emerging issues

Sensitive data privacy is a legal issue, and the detailed legalities are beyond the scope of this paper. Readers are strongly advised to get professional, and localized, legal advice before using medical data. A lot of the law globally is based on the European GDPR but details may well differ – for instance, data privacy usually ends with the data subject's death but in a few areas, deceased subjects have some data privacy rights.

> "Readers are strongly advised to get professional, and localized, legal advice before using medical data."

It is easy to predict the future, but harder to say when it will happen. So, you need to be aware of the future but don't obsess about it to the point of neglecting current reality. For instance:

1 The possible *"death of cryptography"* is something you should be aware of. Quantum computing is coming and may make it easy to break some, widely used, current forms of encryption, but the quantum processing power for this isn't here yet. Since medical records may still be stored 10 years out, when quantum computing will likely be routine, it is good practice now to use ***"quantum safe"*** cryptography for privacy, but you can't really assess the future risk accurately.

2 Similarly, AI is starting to make de-anonymization of anonymized data easy, based on *"quasi identifying data"* (even without AI, if you have data for a whole town and you know that only one person in the town travels by helicopter and has a heliport, say, it might be quite easy to determine which anonymized data refers to that one person). As AI improves or becomes more available, defensible anonymization will become harder.

3 Laws change and the level of enforcement varies. Could vexatious inquiries and requests (perhaps based on *"the right to be forgotten"* in many data privacy codes) be used as the basis for denial-of-service attacks?

4 Emerging healthcare datatypes, such as medical images (DICOM) and DNA data, may bring their own processing challenges.

What this all means is that the treatment of sensitive healthcare information must be flexible and must be reviewed regularly. As technology evolves, what was adequate last year may not be adequate next year. As legal frameworks and public awareness change, risk profiles change. Any technology you adopt for securely managing healthcare information must not compromise flexibility.

AI Analytics

Actor
SE Asian healthcare platform

Preconditions
Entering a crowded marketplace with a new product

Documented as
Product marketing material and website

Description
This platform processes hospital patient data internationally, including in the USA, in a variety of formats, including HL7, X12, and PDFs. It integrates a data searching/anonymizations tool via an integrated API and places emphasis on the safe use of patient data for AI-based analytics. It is intending to add a load balancing feature for the horizontal scaling of large midnight workflows. Effective provision of safe (anonymized) AI-based analytics, with a high-performance capability, is an important selling point for the platform. This use case shows how effective and flexible data anonymizing can facilitate innovation.

What success looks like
Proper risk analysis of the analytics offering, possibly in the context of a trial installation.

Ultimate outcome
Positive, platform selected for purchase or shortlist; negative product not selected.

Personal productivity

Actor
UK NHS

Preconditions
Organizations with devolved or informal, more agile, management

Documented as
Site audits

Description
Data masking tools can be incorporated directly into spreadsheets such as Excel to mask PHI. In general, data masking can help healthcare providers, researchers, and business associates who collect PHI protect it from improper use and disclosure. Using the right data masking or anonymization tools and techniques can enable secure, compliant access to data in a wide range of formats for a wide range of operational needs – from development to analytics to training.

What success looks like
Enablement of effective, efficient end user *"personal productivity"* computing.

Ultimate outcome
Positive, more effective and wider use of IT; negative: failure of end-user computing compliance.

HIPPA data de-identification

Actor
Healthcare providers in the USA

Preconditions
HIPAA healthcare regulations in the USA

Documented as
Compliance reporting

Description
The 1996 US Health Insurance Portability and Accountability Act (HIPAA) requires the de-identification of 18 unique patient attributes, called key identifiers. This is a requirement of the HIPAA Safe Harbor Security Rule, which does not distinguish between data in production or test environments. Healthcare organizations rely on analytics to improve patient care, reduce costs, and streamline operations. However, when database application developers need a realistic test schema or data scientists need to build dashboards or run machine learning models, the PHI in their sources must first be masked. Using unmasked patient data in these environments can lead to data breaches and privacy law violations. Data masking tools allow healthcare entities and business associates to classify, discover, and de-identify PHI in on-premises and cloud databases and file stores. By using deterministic masking functions like format-preserving encryption or unique, consistent pseudonym replacement values, these tools can also preserve referential integrity in masked environments across structured, semi-structured, and unstructured targets, which is necessary for the provision of realistic test coverage.

What success looks like
Third-party, external audit of HIPPA compliance.

Ultimate outcome
Positive, organization is HIPPA compliant; negative, organization fails compliance and incurs various sanctions as well as reputation loss.

Training data

Actor
Medical schools, training centers, and hospitals

Preconditions
Any organization which trains healthcare staff

Documented as
Staff skills catalog and performance

Description
Frequently use case studies, patient histories, and sample datasets for teaching purposes. While data about actual patients is valuable for learning lessons, exposing patient identities is unethical and often illegal. Masking PHI allows educational institutions to provide realistic datasets that reflect actual case complexity and variability without violating privacy laws. By anonymizing quasi-identifying demographic attributes (as discussed in Section 2 above), trainers can share practical examples without risking a data breach or HIPAA violation.

What success looks like
An effective, properly managed training program.

Ultimate outcome
Positive, well-trained healthcare staff with practical experience; negative healthcare staff unaware of the possible complexity of the *"real world"* product not selected.

Summary

To summarise this paper, the vast amount of healthcare data accumulated by healthcare organisations provides them with real opportunities for metrics-based management of their services and for sharing patient data with organizations developing new healthcare technologies.

However, the healthcare sector is highly regulated and, in general, the identity of the subjects of healthcare data must be protected. Moreover, the level of technology adoption in the healthcare industry varies widely, so the chances of meeting older data formats, in the process of being replaced, is high. The chance of meeting global data from outside of your normal sphere of operations is also high.

What this means in practice is that you should partner with data extraction and anonymisation experts, and tool vendors, who have the experience needed to help you mine your data resources and anonymize the data to meet regulatory requirements. Yes, doing this entirely in-house is possible, but many companies don't have the inhouse expertise to do this effectively – the risk is that you fail compliance with critical regulations or, alternatively, that you play it too safe and miss an opportunity.

About the authors

DAVID NORFOLK (LEAD)
Practice Leader:
Development & Governance

David Norfolk was working in the Research School of Chemistry at the Australian National University in the 1970s, when he discovered that computers could deliver misleading answers, even when programmed by very clever people. His ongoing interest in getting computers to deliver useful automation culminated in his joining Bloor in 2007 and taking on the development brief.

Development here refers to developing automated business outcomes, not just coding. It also covers the processes behind automation and the people issues associated with implementing it. He sees organisational maturity as a prerequisite for implementing effective (measured) process automation and ITIL as a useful framework for automated service delivery. He also looks after Collaboration and Business Process Management for Bloor and takes a lively interest in the reinvention of the Mainframe as an Enterprise Server.

David has an honours degree in Chemistry, a graduate qualification in Computing, and was a Chartered IT Professional. He has a somewhat rusty NetWare 5 CNE certification and is a Member of the British Computer Society (he is on the committee of its System Management IT Asset Management Specialist Group). He also has an MA in Visual Communication.

He has worked in database administration (DBA) and operations research for the Australian Public Service in Canberra. David then worked for Bank of America and Swiss Bank Corporation in the UK, holding positions in DBA, systems development method and standards, internal control, network management, technology risk and even PC support. He was instrumental in introducing a formal systems development process for the Bank of America Global Banking product in Croydon.

In 1992 he started a new career as a professional writer and analyst. He is a past co-editor/co-owner) of Application Development Advisor and was associate editor for the launch of Register Developer. He helped organise the first London CMMI Made Practical conference in 2005 and has written for most of the major computer industry publications. He runs his own company, David Rhys Enterprises Ltd, from his home in Chippenham, where he also indulges a keen interest in photography (he holds a Royal Photographic Society ARPS distinction).

PAUL BEVAN (CO-AUTHOR)
Navigator, Research Director:
IT Infrastructure

Paul has had a 40-year career in industry that started in logistics with a variety of operational management roles. For the last 33 years he has worked in the IT industry, mostly in sales and marketing, covering everything from mainframes to personal computers, development tools to specific industry applications, IT services and outsourcing. In the last few years he has been a keen commentator and analyst of the data centre and cloud world. Until recently he was also a non-executive director in an NHS Clinical Commissioning Group.

Paul has a deep knowledge and understanding about the IT services market and is particularly interested in the impact of Cloud, Software Defined infrastructure, OpenStack, the Open Compute Project and new data centre models on both business users and IT vendors. His mix of business and IT experience, allied to a passionate belief in customer focus and "grown-up" marketing, has given him a particular capability in understanding and articulating the business benefits of technology. This enables him to advise businesses on the impact and benefits of particular technologies and services, and to help IT vendors position and promote their offerings more effectively.

Bloor overview

Technology is enabling rapid business evolution. The opportunities are immense but if you do not adapt then you will not survive. So in the age of Mutable business Evolution is Essential to your success.

We'll show you the future and help you deliver it.

Bloor brings fresh technological thinking to help you navigate complex business situations, converting challenges into new opportunities for real growth, profitability and impact.

We provide actionable strategic insight through our innovative independent technology research, advisory and consulting services. We assist companies throughout their transformation journeys to stay relevant, bringing fresh thinking to complex business situations and turning challenges into new opportunities for real growth and profitability.

For over 25 years, Bloor has assisted companies to intelligently evolve: by embracing technology to adjust their strategies and achieve the best possible outcomes. At Bloor, we will help you challenge assumptions to consistently improve and succeed.

Copyright and disclaimer

BLOOR RESEARCH

20–22 Wenlock Road, London N1 7GU, UK

+44 (0) 1494 311 460

info@Bloorresearch.com

www.Bloorresearch.com